Splitwise: Efficient Generative LLM Inference Using Phase Splitting

Year: 2023 | Citations: 408 | Authors: Pratyush Patel, Esha Choukse, Chaojie Zhang, Íñigo Goiri, Aashaka Shah

---

Abstract

Generative large language model (LLM) applications are growing rapidly, leading to large-scale deployments of expensive and power-hungry GPUs. Our characterization of LLM inference shows that each inference request undergoes two phases: a compute-intensive prompt computation phase and a memory intensive token generation phase, each with distinct latency, throughput, memory, and power characteristics. Despite state-of-the-art batching and scheduling, the token generation phase underutilizes compute resources. Unlike prompt computation, token generation does not need the compute capability of the latest GPUs and can be run with lower power and cost. Based on these insights, we propose Splitwise, a model deployment and scheduling technique that splits the two phases of LLM inference requests on to separate machines. Splitwise enables phase-specific resource management using hardware that is well suited for each phase. Request state is transferred efficiently between machines using optimized network libraries on the fast back-plane interconnects available in today's GPU clusters. Using Splitwise, we design homogeneous and heterogeneous LLM inference clusters optimized for throughput, cost, and power Compared to current designs, Splitwise clusters achieve up to $1.4 \times$ higher throughput at $\mathbf{2 0 \%}$ lower cost. Alternatively, they can deliver $2.35 \times$ more throughput under the same power and cost budgets.